知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

基于专利新颖性研发组合标识图的技术机会分析

——以溶栓药领域为例

◎ 陶成琳 王伟 张世玉

吉林大学公共卫生学院医学信息学系 长春 130021

摘要：[目的 / 意义] 识别新颖专利代替由一组关键词代表的专利空白，改善技术机会识别过于主观的缺陷。[方法 / 过程] 利用一种基于系统流程的定量方法识别专利的新颖程度。通过基于密度的局部离群点（DLOF）算法识别出新颖专利，利用技术范围指标与同类专利数量指标构建研发组合标识图。[结果 / 结论] 研究结果表明，基于专利的新颖性研发组合标识图可以准确地识别出新颖专利，为技术研发提供借鉴。

关键词：技术机会分析 新颖专利 专利新颖性研发组合标识图 文本挖掘 局部异常因子

分类号：G202

引用格式：陶成琳，王伟，张世玉．基于专利新颖性研发组合标识图的技术机会分析——以溶栓药领域为例 [J/OL]. 知识管理论坛 , 2016, 1(4): 276-282[引用日期]. http://www.kmf.ac.cn/p/1/44/.

1 引言

随着新技术的不断发展，企业逐渐意识到技术创新本身固有的风险。在这样的背景下，技术机会分析的战略重要性得到进一步提升。越来越多的企业开始通过组织专家小组，掌握产业的技术发展现状，力求找到技术突破口。随着技术的不断增值，劳动密集型产业被技术创新型产业逐步取代，耗费人力及时间的专家讨论已经不能满足创新周期逐步缩短的产业布局 [1]。因此，企业领导者希望通过技术机会分析发现潜在的技术机会，节约生产成本及时间创造更多价值。基于文献计量分析与文本挖掘的专利分析可以用来分析技术机会 [2]，客观的识别出企业的技术发展点。

基于专利的技术机会分析具体是指利用一系列技术以及大规模数据集对专利信息进行挖掘，发现新技术及预测其市场前景，偏重于技术预测和预见。由于专利信息包含技术领域的

基金项目：本文系吉林省教育厅“十二五”社会科学研究规划课题“基于专利分析的吉林省生物制药产业创新发展战略与实施路径研究”（项目编号：2014B019）的研究成果。
作者简介：陶成琳（0000-0003-3611-6075），硕士研究生；王伟，教授，博士生导师，通讯作者，E-mail：w_w@jlu.edu.cn；张世玉（ORCID：0000-0002-6320-9982），博士研究生。
收稿日期：2016-06-13 发表日期：2016-08-25 本文责任编辑：易飞

知识管理论坛，2016(4):276-282
DOI：10.13266/j.issn.2095-5472.2016.033

关键性技术，对专利信息进行分析可以为技术机会提供支持。技术机会分析的结果可以利用图形、表格、曲线图等形式的专利地图表示，使结果更加简洁、直观[3]。目前已有诸多研究将文本挖掘与数据降维等数据挖掘方法相结合，对专利进行技术机会分析。同时，专利空白地图[4]、专利空缺地图[5]、基于生成式拓扑映射的专利地图及语义专利地图[6]等方法都被用来进行技术机会的识别。虽然许多学者都从不同角度对技术机会的识别进行研究，但仍不够精确。现阶段的研究多是将高维的专利数据根据其技术之间的相似性实现数据降维，使其显示在一个二维专利地图中。专利地图中的空白点被解释为潜在技术机会，空白点的密度越低，则潜在的技术机会就越大。这样的研究虽然可以节省人工操作的时间，但在实践中分析大量的非结构化数据时依然存在弊端。目前关于专利地图的研究侧重于可视化的方法[7]，其中最主要的是主成分分析（PCA）与自组织映射（SOM）方法。主成分分析法[8]是通过线性旋转变换寻找方差最大的方向作为坐标轴方向，舍弃方差较小的维度实现降维，但是舍弃的变量中所包含的有价值的信息并没有被考虑在内。自组织映射[9]方法可以实现向量的非线性映射和聚类，在二维平面上进行专利数据的预测，具有良好的效果。但其认为相邻输入向量彼此无关，未考虑到专利信息是一个时间序列，相邻时刻的数据之间具有较大的相关性，因此采用自组织映射方法处理专利问题会丢失很多信息。关于技术机会分析，最新的研究应用生成式拓扑映射[10]识别专利空白点，利用逆映射保持原拓扑关系不变生成专利地图识别出一种由关键词组成的关键词串，这组关键词串是否有意义以及判断专利空白的阈值都是依赖专家小组的意见，使结果过于主观[11]。由于专利空白评估存在一定程度的主观性，因此对潜在技术机会的评价也存在一定的主观性，目前尚缺乏衡量专利新颖程度的客观指标。

鉴于以上存在的弊端，利用一种改进的方法识别和评价技术机会是必要的。本文基于系统流程与定量的方法识别出新型专利，取代之前的一组代表专利的关键词的专利空白，使潜在技术机会的识别更加精确。所提出的方法的核心是文本挖掘与局部异常因子，利用文本挖掘的方法提取专利关键词之后采用局部异常因子衡量一个数据集合的新颖程度。与现阶段的以知识流动[12]与知识链接[13]为基础进行技术预测的方法不同，是以关键词使用的相似程度即专利信息的新颖程度为衡量指标。区别于现有研究，应用技术范围指标取代被引频次指标，与同类专利数量指标构建二维研发组合标识图，使专利标识图简洁精确的同时更加适用于分析中文文本的专利信息。

2 新颖专利研发组合标识图的构建

2.1 构建专利矩阵

传统的专利矩阵，将一个系统分解为几个相互排斥且完全穷尽的二维矩阵。这种方法允许系统对各维度进行单独处理，汇总后形成对整个数据的分析。本文对专利的文本信息进行挖掘，将专利的签发日期、专利号以及专利关键词所构建的以关键词（S_{11}，…，S_{nn}）为横轴、专利号（P_1，…，P_n）为纵轴的专利矩阵作为局部离群点检测的输入值。表 1 中列举的专利矩阵中，签发日期与专利号用文本形式表示，而关键词向量则用二进制值表示，“1”代表该专利与列出的关键词相关，而“0”则意味着不相关。表一中，P_5 是在 YMD_5 发表的，在 D_1 的维度具有 S_{1i} 关键词，D_2 有 S_{21} 关键词。具体形态如表 1 所示：

2.2 识别新颖专利

基于密度的局部异常挖掘算法不再把异常看做是一种二元属性 (只有异常与非异常的区别)，而是用局部异常因子 LOF 来表示对象的异常程度。对象 p 的局部异常因子反映了该对象的异常程度，局部异常因子 $LOF_k(p)$ 越大，则该对象是异常数据的可能性越大；反之，则该对象是异常数据的可能性越小。

KMF www.kmf.ac.cn 知识管理论坛 2016 年 第 4 期（总第 4 期）

表 1 专利矩阵示例

专利号	公布日期	D_1		D_2	…	D_n	
		S_{11} …	S_{1i}	S_{21} …	S_{2j}	S_{n1} …	S_{nm}
P_1	YMD_1	1 …	0	0 …	0	0 …	0
P_2	YMD_2	1 …	0	1 …	0	0 …	0
P_3	YMD_3	1 …	0	0 …	0	0 …	0
P_4	YMD_4	0 …	0	1 …	1	1 …	0
P_5	YMD_5	0 …	1	1 …	0	0 …	0
P_6	YMD_6	0 …	0	0 …	0	0 …	1
P_7	YMD_7	0 …	0	0 …	0	0 …	1
P_8	YMD_8	0 …	0	1 …	1	1 …	1

本文所提出的方法利用基于密度的局部异常因子（LOF) 评估专利的异常程度，以实现定量结果的客观解释。该方法可以在过滤异常值的基础上检测任意形状的自然聚类。某一点的局部异常值通过该点与周围各点的平均密度比值得到。具体计算分为 4 个步骤：①对象 p 的 k- 距离 (k-distance(p) 为 p 与其近邻 k 的欧几里得距离，其中 k 值被定义为参数聚类的最小距离；②将 q 定义为 p 的可达距离，表示为 reachDistk(p,q)，通过 max{d(p,q), k-distance(p)} 得到，d(p,q) 即 p 与 q 之间的欧几里得距离；③ $N_k(p)$ 定义为 p 的 k 近邻点的集合，以密度为基础的可达距离表示为 $lrd_k(p)$，如公式（1）所示；④ p 对其周围对象 k 的 LOF 值如公式（2）所示。

$$lrd_k(p)=\frac{k}{\sum_{q\in N_k(p)} reachDist_k(p,q)} \quad (1)$$

$$LOF(p)=\frac{1}{k}\sum_{q\in N_k(p)}\frac{lrd_k(q)}{lrd_k(p)} \quad (2)$$

评估专利新颖程度的过程主要分为两步：

（1）LOF 的计算。假设 PS_j 被定义为一组 j 年发布的专利，其中每个专利在形态学矩阵中都由关键词向量表示（S_{11}，S_{12}，…S_{nm}）。对于专利 P_i，我们将计算出的 LOF_j（P_i）定义为专利 P_i 的局部异常值。在本文所提出的方法中，k 值被定义为专利的数量，如大多数形态结构专利数据一样，将 k 值定义为专利数量。通过确定合适的 k 值，可以计算出所有专利的 LOF 值。

（2）LOF 的标准化。通过 LOF 值我们可以分析出某年度专利的新颖程度。但由于某一专利的 LOF 值在不同年份存在差异，随着时间推移，LOF 值的变化较难掌握。即使某专利在不同的两年中 LOF 的值相同，其新专利也会因不同专利集合而有所差异。为解决这一问题，我们引入核密度估计的方法使 LOF 的值标准化。核密度估计是从离散样本中确定概率分布函数的非参数估计方法。J 年发表专利的概率分布函数定义为 f_j(LOF)，通过核密度估计计算，如公式（3）所示：

$$f_j(LOF)=\frac{1}{n(PS_j)h}\sum_{i=1}^{n(PS_j)}K[\frac{LOF-LOF_j(P_i)}{h}] \quad (3)$$

对于 $LOF_j(P_i)$，当 i=1,…,n 时为样本点，$n(PS_j)$ 是专利的数量，K 为高斯核函数，h 为平滑因子，计算出的 $R_j(P_i)$ 为每个 $LOF_j(P_i)$ 的相对新颖率，如公式（4）所示：

$$R_j(P_i)=F(LOF(p_i))=\int_{-\infty}^{LOF(P_i)} f_j(LOF)dLOF \quad (4)$$

相对新颖率 $R_j(P_i)$ 与 j 年公布的专利 P_i 相比 LOF 值较低，因此我们可以将 P_i 作为 R_j（P_i）新专利的扩大值。其允许对某一年专利的 LOF 值进行比较以及对特定的专利新颖性进行动态分析。

知识管理论坛，2016(4):276-282
DOI：10.13266/j.issn.2095-5472.2016.033

2.3 构建新颖专利研发组合标识图

在识别出新颖专利的基础上，本文利用专利技术范围与同类专利数量这两个指标，综合分析并识别专利技术机会。现阶段的研究应用专利被引频次评估技术机会，认为专利的被引频次越高，该专利拥有的技术与经济影响力就越大，但专利的审批与专利发布之间存在时间差。越来越多的研究指出，较早公开的专利与新专利的被引频次在一定程度上受到时间的制约，因此本文应用技术范围这一指标替换专利被引频次这一维度。专利技术范围是指某一专利的分类号的数量及范围，每个专利除去其拥有的主分类号，其分类号数量越多说明该技术覆盖的范围越广，影响力也越大。而随着知识产权保护的发展，专利侵权出现的频率逐年上升，所带来的经济赔偿也逐年提高。受收益率与专利价值的制约，同类专利的数量一定程度上可以说明专利的可效仿程度，同类专利数量越少，该专利可效仿程度越高。基于这些因素，新颖专利研发组合标识图的构建以技术范围与同族专利的数量为评价指标。

3 实证研究——以溶栓药领域为例

近年来，随着人口老龄化的加剧，外周血管疾病的发病率呈逐年上升趋势。在美国，60岁以上的人群中有多达 5% 的男性和 2.5% 的女性患者存在间歇性跛行症状。溶栓药、抗血小板药是降低心脑血管病发生几率的有效药物 [14]。医药相关企业着力于研制新型、高效溶栓药物，因此，在该领域识别出新的技术机会是必要的。

3.1 数据来源与方法

本研究的数据来源于中国知网（CNKI）的中国专利数据库，检索截止到 2015 年 11 月 15 日溶栓药领域的全部发明专利和实用新型专利。共获得 561 个相关专利族，去重后利用 Microsoft Office Access 建立包含标题、申请人、地址、公布日期、公开号等条目的专利数据库，并基于专利号、公布日期与摘要构建专利矩阵。

3.2 构建专利矩阵

利用文本挖掘软件 TextAnalysis 2.1 基于 TF-IDF 指数发现描述溶栓药物特征的 8 个维度的 25 个重要关键词，再进一步通过 Salton 索尔顿海指数找到关键词的共现关系，使关键词的确定更加精确高效。

溶栓药领域的 8 个维度分别为：制备方法、医疗器械、作用部位、中药组合物、医药组合物、治疗方法、基因表达与生物制药。具体矩阵见表 2、表 3。

表 2 溶栓药物的专利矩阵结构

维度	具体描述	关键词
制备方法（D1）	提取（S11）	丹参、三七 …
	纤溶酶（S12）	突变、赖氨酸 …
	重组（S13）	蛋白、DNA 序列 …
医疗器械（D2）	导管（S21）	逆向溶栓、药物注射 …
	超声（S22）	破菌处理、空化效应 …
作用部位（D3）	腔静脉（S31）	闭塞、抗凝 …
	深静脉（S32）	融合蛋白、脂质体 …
中药组合物（D4）	成分组合（S41）	田七、川穹、丹参 …
生物制药（D5）	融合蛋白（S51）	凝胶层析、亲和层析 …
	蛋白质药物（S52）	人组织型、大肠杆菌 …
治疗方法（D6）	渗透能力（S61）	超声、无空化 …
	支架（S62）	疏水、水解 …
	ELISA(S63)	检测、溶栓素蛋白 …
基因表达（D7）	载体（S71）	重组、重组菌 …
	溶栓酶（S72）	沙蚕激酶、表达系统 …



表 3 部分溶栓药物的专利矩阵

专利号	公开日期	制备方法			…	基因表达	
		提取	纤溶酶	配方	…	载体	溶栓酶
CN1037275	19891122	0	0	1	…	0	0
CN1037345	19891122	0	1	0	…	0	0
CN1053007	19910717	0	0	1	…	0	0
…	…	…	…	…	…	…	…
CN105012381A	20151104	0	0	1	…	0	0
CN105039377A	20151111	0	0	0	…	1	1
CN105072914A	20151118	0	0	0	…	0	0

3.3 识别新颖专利

通过专家讨论确定主要专利的数量后，k 值的确定分为两个步骤，首先通过计算确定多数专利的数量，然后测量专利矩阵中关键字向量之间的余弦相似性，帮助专家判断具体的 k 值。余弦相似度被认为是计算两个非结构化文档之间相似度的最常用指标，计算方法如公示（5）所示：

$$\cos\theta = \frac{A \bullet B}{|A||B|} \qquad (5)$$

A 与 B 即代表文档中的关键词向量。相似性的范围定义为 0 至 1，两个文件相似性越高，则值也越大。通过计算最终确定 k 的值为 10，通过 Metlab 计算 LOF 值及标准化的 LOF 值，得出每个专利的新颖程度评估数值，部分专利相对新颖程度如表 4 所示：

表 4 部分溶栓药物的专利新颖程度

专利号	公开日期	专利新颖程度					
		2010	2011	2012	2013	2014	2015
CN1037275	19891122	0.0451	0.0333	0.0046	0.0021	0.0046	0.0083
CN1037345	19891122	0.0394	0.0474	0.0342	0.0659	0.0097	0.0057
CN1053007	19910717	0.0392	0.0213	0.0237	0.0096	0.0073	0.0049
…	…	…	…	…	…	…	…
CN105012381A	20151104	–	–	–	–	–	0.3836
CN105039377A	20151111	–	–	–	–	–	0.2453
CN105072914A	20151118	–	–	–	–	–	0.5352

3.4 新颖专利研发组合标识

排名前 5% 的专利研发组合标识图如图 1 所示，根据专利矩阵的信息将专利用圆圈表示，圆圈的大小表示专利的重要程度。横轴与纵轴的分类线分别为专利技术范围与同类专利数量经过标准化计算后的平均值。专利应用数量的高低也决定了专利影响力的大小，而专利的价值及潜在价值需要通过调查确定。通过本文所构建的方法识别出专利 CN102311396A 与 CN102210666A 为溶栓药领域的技术机会。吡嗪类衍生物的抗氧化作用和溶栓作用，制备治疗由于自由基过量产生或血栓引起的心、脑血管系统疾病以及退行性老化疾病等的新型药物。丹酚酸 A 可用于制备促进血或血浆中

知识管理论坛，2016(4):276–282
DOI：10.13266/j.issn.2095–5472.2016.033

CAMP 含量升高的药物，制备抑制血小板中磷酸二酯酶活性的药物，特别是可用于制备预防或治疗溶栓、经皮冠状动脉介入术或冠脉搭桥术等原因引起的心肌缺血再灌注损伤的药物。相关研发企业可就以上两个专利相关技术进行重点研发。

本文所构建的新颖专利研发组合标识图是对特定时间内某一横截面的评价，识别出的结果会随着技术的发展不断变化，如果特定的技术潜在机会进入一个新的阶段则其在标识图中的信息需要被移除。虽然专利研发组合标识图的结构是不固定的，但初始的专利矩阵建立后是可以重复使用的，我们只需要将新生产出的专利信息加入到矩阵中即可。

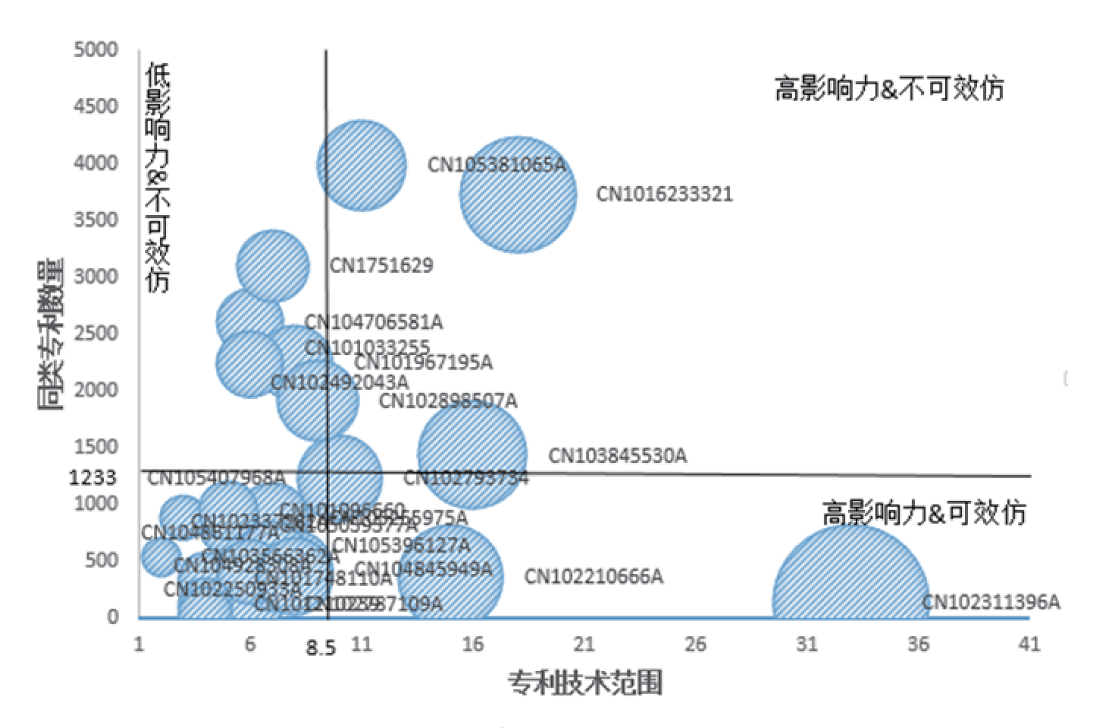


图 1 新颖专利研发组合标识图

4 结论

本研究基于系统流程与科学方法提出了一种利用识别新颖专利进行技术机会识别的方法：通过确定异常专利，而不是由一组关键词组成的专利关键词新组合，使技术机会的识别更加精确。另外，专利新颖程度可以通过量化指标进行比较，使结果更加客观。技术机会分析战略被认为是新兴技术生成竞争情报的有效手段 [15]。本文基于定量数据与系统流程提出了一种技术机会识别的方法，将基于密度的局部离群点（LOF）的应用领域从过程控制、故障检测扩展为通过文本挖掘技术对新型专利信息进行整合和解释。与当前的研究不同，应用技术范围取代被引频次这一指标，与同类专利数量构建二维研发组合标识图，使专利标识图更加简洁精确。本文所提出的方法的局限性在于，所生成标识图中出现的并不是一套明确的新专利，而是将可能性极高的新专利识别出来，且在一定程度上仍然要依靠专家评价确定。今后的改进方向为逐步提高量化指标衡量的程度，减少主观评价从而进一步优化技术机会识别流程。

参考文献：

[1] LEE C, SONG B, PARK Y. How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships[J]. Technology analysis & strategic management, 2013, 25(1): 23-38.

[2] LEE C, JEON J, PARK Y. Monitoring trends of technological changes based on the dynamic patent lattice: a modified formal concept analysis approach[J]. Technological forecasting & social change, 2011, 78(4): 690-702.
[3] YOON B. Strategic visualisation tools for managing technological information[J]. Technology analysis & strategic management, 2010, 22(3): 377-397.
[4] YOON B, YOON C, PARK Y. On the development and application of a self-organizing feature map-based patent map[J]. R & D management, 2002, 32(4): 291-300.
[5] LEE S, YOON B, PARK Y. An approach to discovering new technology opportunities: keyword-based patent map approach[J]. Technovation, 2009, 29(6/7): 481-497.
[6] Isumo Bergmann , Daniel Butzke , Walter L, et al. Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips[J]. R & D management, 2008, 38(5): 550-562.
[7] LEE C, SON C, YOON B, et al. An instrument for discovering new mobile service opportunities[J]. International journal of mobile communications, 2013, 11(4): 374-392.
[8] 李靖华 , 郭耀煌 . 主成分分析用于多指标评价的方法研究——主成分评价 [J]. 管理工程学报 , 2002, 16(1): 39-43.
[9] 余健 , 郭平 . 自组织映射 (SOM) 聚类算法的研究 [J]. 现代计算机 (专业版), 2007(3): 7-8.
[10] SON C, SUH Y, JEON J, et al. Development of a GTM-based patent map for identifying patent vacuums.[J]. Expert systems with applications, 2012, 39(39):2489-2500.
[11] LEE C, SON C, YOON B, et al. An instrument for discovering new mobile service opportunities[J]. International journal of mobile communications, 2013, 11(4):374-392.
[12] REITZIG M. What determines patent value?[J]. Research policy, 2003, 32(1):13-26.
[13] CUNNINGHAM S W. Analysis for radical design[J]. Technological forecasting & social change, 2009, 76(9): 1138–1149.
[14] 刘健 . 口服抗血小板药物在冠心病治疗中的应用 [J]. 中国循环杂志 , 2012, 27(4): 103-104.
[15] PORTER A L, JIN X Y, GILMOUR J E. Technology opportunities analysis: integrating technology monitoring, forecasting, and assessment with strategic planning[J]. SRA journal, 1994, 26(2):21-31.

作者贡献说明：
陶成琳：确定选题、研究方法和研究领域，处理数据，总结结论撰写论文；
王伟：确定选题方向、研究结构，修改论文并定稿；
张世玉：参与选题及研究领域的确定、核实数据、修订论文。

Technology Opportunity Analysis Based on the Combination of Patent Novelty Research and Development—A Case Study of Thrombolytic Drugs

Tao Chenglin Wang Wei Zhang Shiyu
Department of medical informatics, School of Public Health, Jilin University, Changchun 130021

Abstract: [Purpose/significance] The meanings of potential technology opportunities become more explicit by identifying anomaly patents rather than patent vacancies that are usually represented as a simple set of keywords. **[Methods/Process]** We propose an approach to detecting anomaly patents based on systematic processes and quantitative outcomes. Density-based Local Outlier (DLOF) algorithm is used to identify novelty patents, then use scope of technology index with amount of similar patents index structure Anomaly-portfolio patent map. **[Results/Conclusion]** Research results show that novelty-focused patent mapping for technology opportunity analysis can accurately identify the novelty patents and provide a reference for the technology research and development.

Keywords: technology opportunity analysis novelty-focused novelty-focused patent identification map text mining local outlier factor